PCGC Social Determinants of Health and Geocoding Manual

March 27, 2024

Introduction

The PCGC geocoding docker container provides means of geocoding a given list of PCGC participant addresses and augmenting it with additional location-derived information (geomarkers) as well as determining the driving distance (in minutes) to a PCGC center. After installation, the software runs on a local computer without requiring an internet connection, thus maintaining the security and privacy of the participant information. The underlying software is based on Cole Brokamp's deGAUSS package.

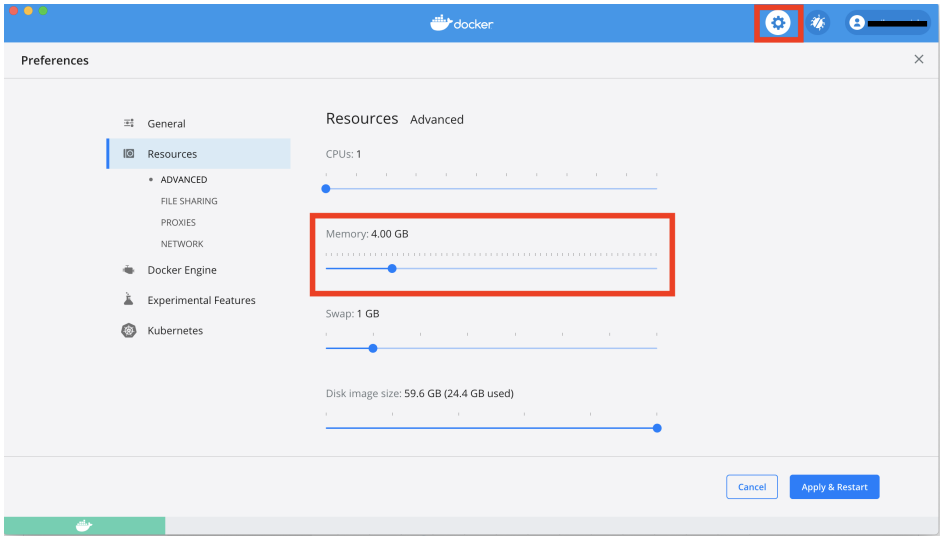
Requirements

- Operating System:
  - MacOS
  - Windows
- RAM: 8GB
- Disk Space: 20GB (docker container is 10GB)
- administrator privileges (initially only, to install the 'docker' software)

Step 0: Install Docker

See the Installing Docker webpage.

Note about Docker Settings:
After installing Docker, but before running containers, go to **Docker Settings > Advanced** and change **memory** to greater than 4000 MB (or 4 GiB)



If you are using a Windows computer, also set **CPUs** to 1.

Step 1: Preparing Your Address File

The address file must be a CSV file with either a column titled `address` containing all address components or columns titled `lat` and `lon` with the participant's latitude and longitude, respectively. Other columns may be present - in particular a `participant ID` column and an `address_date` column are recommended. However, the software will ignore (but preserve) all additional columns besides `address`, `lat` and `lon`.

An example address CSV file might look like the following `address-sample-date-UTAH.csv` file from the docker container:

address-sample-date-UTAH

ID	address_date	address	lat	lon
1	1/1/24		39.98248794	-112.2851437
1	1/1/23		39.99908182	-112.1954693
2			40.62336655	-112.0137647
3	5/6/22	5331 Rexford Court, Montgomery AL 36116		
3	3/3/21	6095 Terry Lane, Golden CO 80403		
4	5/4/23	4016 Doane Street, Fremont CA 94538		

Example address CSV files are my_address_file.csv, address-sample.csv or address-sample-date-UTAH.csv, all located in the tests folder of the docker container source.

Note: Please make sure to enclose the information in the `address` column in quotation marks (e.g., " ") if it contains commas.

Step 2: Running the PCGC container (the short version)

If `my_address_file.csv` is an address file in the current working directory with an address column named `address`, then the command to process it through the PCGC geocoding container is:

- macOS:

```
docker run --rm -v $PWD:/tmp ghcr.io/pcgcid/geocoder_pcgc:0.0.1 \
-s PCGC_UTAH -i my_address-file.csv -o UTAH_output
```

- Windows (CMD):

```
docker run --rm -v "%cd%":/tmp ghcr.io/pcgcid/geocoder_pcgc:0.0.1 ^
-s PCGC_UTAH -i my_address-file.csv -o UTAH_output
```

will produce 3 output files:

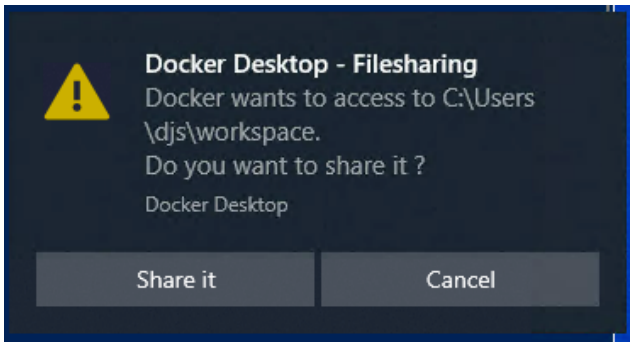
- `UTAH_output.csv`: This file has full output data, **including PII data**. Do NOT sent this to the ACC.
- `UTAH_output-deid.csv`: This file contains de-identified fields specified by the user as well as location-derived information. By default, the list of de-identified fields contain "id","address_date","matched_state","precision", "fraction_high_school_edu","median_income","fraction_no_health_ins","fraction_poverty","fraction_vacant_hous "dep_index","drivetime_selected_center","nearest_center_pcgc","drivetime_pcgc","version". "id" and "address_date" are copied verbatim from the input address file; it is the user's responsibility to ensure they don't contain PHI

- `UTAH_output-log.txt`: This file is an output log of the processing.

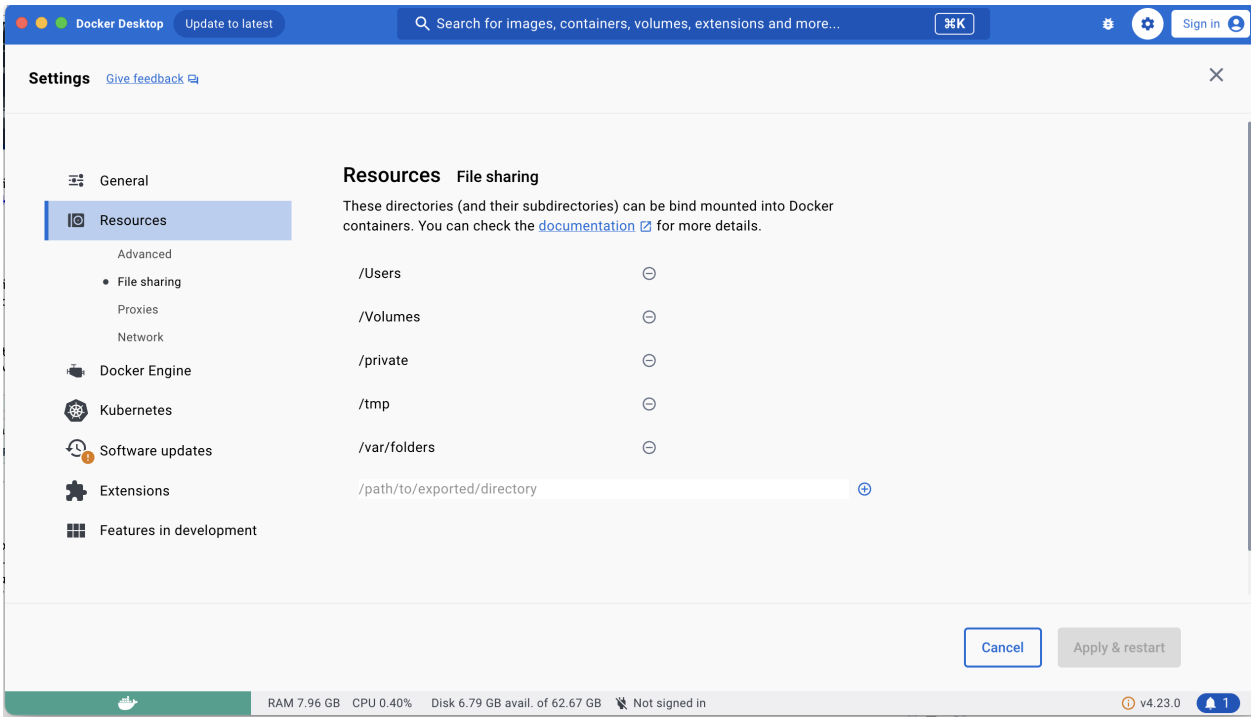
Note: The example above uses `-s PCGC_UTAH` for PCGC center in Utah. To change the care center, replace `-s PCGC_UTAH` with one of the site abbrevations below (e.g., `-s PCGC_YALE`):

Abbreviation	**Name**
PCGC_YALE	`Yale`
PCGC_BOSTON	`Boston Childrens`
PCGC_MTSINAI	`Mt. Sinai`
PCGC_COLUMBIA	`Columbia`
PCGC_CHOP	`CHOP`
PCGC_UTAH	`Utah`
PCGC_CHLA	`Childrens of LA`

Note: On Windows computers you may need to give Docker explicit permissions to access the folder containing the address file (and possibly restart the Docker daemon after you have done so).



However, if notifications are disabled, the confirmation box will not appear and Docker will automatically decline the permission. In that case, go to **Docker Settings > Resources > FileSharing**. Add required folder and hit Apply & Restart



Note: The first time this process is run, docker will download the latest container from the ACC, which takes a few minutes of time. Later runs will not require internet connections (unless the container is to be updated with the latest version).

Note: After processing, **please** inspect the output files and fix obvious formatting problems with the address file should they arise (see also the section below on input address data formatting). The `*-deid.csv` file is safe to be sent to the ACC via secure upload to AWS (similar to the EMR data uploads).

Running the PCGC deGAUSS container (the longer version)

Command line parameters to show help, version and site list are as follows:

- `-h` or `--help`: Show available parameters. For example, users can use this command:

  ```
  docker run ghcr.io/pcgcid/geocoder_pcgc:0.0.1 -h
  ```

  or

  ```
  docker run ghcr.io/pcgcid/geocoder_pcgc:0.0.1 --help
  ```

- `-v` or `--version`: Show the current version of Docker container with this command:

  ```
  docker run ghcr.io/pcgcid/geocoder_pcgc:0.0.1 -v
  ```

  or

  ```
  docker run ghcr.io/pcgcid/geocoder_pcgc:0.0.1 --version
  ```

- `--site-list`: Print all available sites with this command:

```
docker run ghcr.io/pcgcid/geocoder_pcgc:0.0.1 --site-list
```

This container **requires** both of the following arguments:

- `-i` to specify the path to the input address CSV file
- `-s` or `--site` to specify theabbreviation of the PCGC center of interest

Abbreviation	Name
PCGC_YALE	`Yale`
PCGC_BOSTON	`Boston Childrens`
PCGC_MTSINAI	`Mt. Sinai`
PCGC_COLUMBIA	`Columbia`
PCGC_CHOP	`CHOP`
PCGC_UTAH	`Utah`
PCGC_CHLA	`Childrens of LA`

This container takes the following optional arguments:

- `-o` or `--output-file-prefix` to specify prefix of output files. By default, the prefix is `output`, which will generate output.log, output-phi.csv, output-deid.csv
- `--f` or `--include-deid-fields` to specify list of fields to include in output. Default fields:
  - id, address_date, precision, geocode_result, fraction_assisted_income, fraction_high_school_edu, median_income, fraction_no_health_ins, fraction_poverty, fraction_vacant_housing, dep_index, drivetime_selected_center, nearest_center_pcgc, drivetime_pcgc, version
- `--force` to force the container to overwrite output files if one of the output files already exists. By default, the program would exit if one of the output files already exists

Running the PCGC container (additional details)

This Docker image does the following:

1. perform geocoding on addresses (if not geocoded already, i.e., if `lat` and `lon` are not specified in the input), adding the following columns:

- **`matched_street`**, **`matched_city`**, **`matched_state`**, **`matched_zip`**: matched address componets (e.g., `matched_street` is the street the geocoder matched with the input address); can be used to investigate input address misspellings, typos, etc.
- **`precision`**: The method/precision of the geocode. The value will be one of:
  - `range`: interpolated based on address ranges from street segments
  - `street`: center of the matched street
  - `intersection`: intersection of two streets
  - `zip`: centroid of the matched zip code
  - `city`: centroid of the matched city
- **`score`**: The percentage of text match between the given address and the geocoded result, expressed as a number between 0 and 1. A higher score indicates a closer match. Note that each score is relative within a precision method (i.e. a `score` of `0.8` with a `precision` of `range`is not the same as a `score` of `0.8` with a `precision` of `street`).

- `lat` and `lon`: geocoded coordinates for matched address
- `geocode_result`: A character string summarizing the geocoding result. The value will be one of
  - `geocoded`: the address was geocoded with a `precision` of either `range` or `street` and a `score` of `0.5` or greater.
  - `imprecise_geocode`: the address was geocoded, but results were suppressed because the `precision` was `intersection`, `zip`, or `city` and/or the `score` was less than `0.5`.
  - `po_box`: the address was not geocoded because it is a PO Box
  - `non_address_text`: the address was not geocoded because it was blank or listed as "foreign", "verify", or "unknown"
- then join with tract-level deprivation index data derived from the 2018 American Community Survey (ACS), adding the following columns:
  - `fips_tract_id`: 2010 census tract identifier
  - 2018 American Community Survey variables:
    - `fraction_assisted_income`: fraction of households receiving public assistance income or food stamps or SNAP in the past 12 months
    - `fraction_high_school_edu`: fraction of population 25 and older with educational attainment of at least high school graduation (includes GED equivalency)
    - `median_income`: median household income in the past 12 months in 2018 inflation-adjusted dollars
    - `fraction_no_health_ins`: fraction of population with no health insurance coverage
    - `fraction_poverty`: fraction of population with income in past 12 months below poverty level
    - `fraction_vacant_housing`: fraction of houses that are vacant
  - `dep_index`: composite measure of the 6 variables above
- then compute drive time to a Pediatric Cardiac Genomics Consortium (PCGC) specified by user, adding the following columns:
  - `drivetime_selected_center`: computed estimated drive time to center specified by user
  - `nearest_center_pcgc`: Nearest PCGC center as computed by the Docker image
  - `distance_pcgc`: Distance to the nearest PCGC center as computed by the Docker image

Details on the processing steps contained in the software

1. Geocoding

Input address data formatting

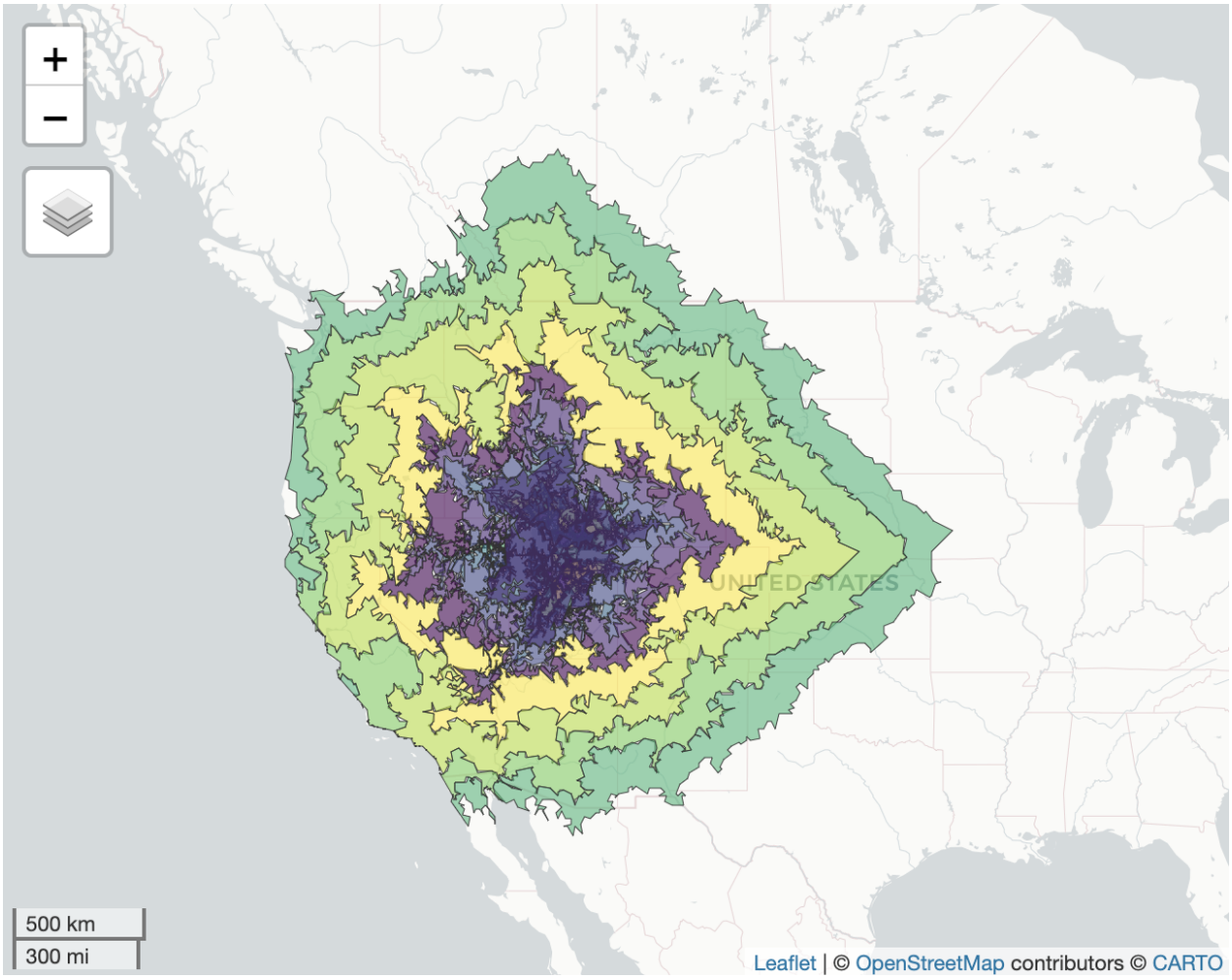
- Other columns may be present, but it is recommended to only include `address`, an optional identifier column (e.g., `id`) and an optional `address_date` column.
- Address data must be in one column called `address`.
- Separate the different address components with a space
- Do not include apartment numbers or "second address line" (but its okay if you can't remove them)
- ZIP codes must be five digits (i.e. `32709`) and not "plus four" (i.e. `32709-0000`)
- Do not try to geocode addresses without a valid 5 digit zip code; this is used by the geocoder to complete its initial searches and, if missing, will likely return incorrect matches
- Spelling should be as accurate as possible, but the program does complete "fuzzy matching" so an exact match is not striclty necessary
- Capitalization does not affect results

- Abbreviations may be used (i.e. `St.` instead of `Street` or `OH` instead of `Ohio`)
- Use Arabic numerals instead of written numbers (i.e. `13` instead of `thirteen`)
- Address strings with out of order items could return NA (i.e. `3333 Burnet Ave Cincinnati 45229 OH`)
- Geomarker data used was prepared following the instructions here using the 2021 TIGER/Line Street Range Address files from the Census

2. Deprivation index This container overlays the input latitude and longitude coordinates with 2010 census tracts, then joins with tract-level deprivation index data derived from the 2018 American Community Survey (ACS).

For more information on the deprivation index, please see the deprivation index page.

3. Drive time This container uses isochrones to assign drive time to care center for each input address. Drive time isochrones are concentric polygons, in which each point inside a polygon has (roughly) the same drive time to the care center. Below is an example of drive time isochrones around the PCGC center in Utah



Drive time isochrones were obtained using a self-hosted openroute service in order to overcome the time limitations of the publicly available API.

We defined 24 levels of isochrones with driving distances up to 960 minutes (16 hours): 15 30 45 60 75 90 105 120 135 150 165 180 195 210 225 240 300 360 420 480 600 720 840 960 minutes

DeGAUSS Details

For detailed documentation on DeGAUSS, including general usage and installation, please see the DeGAUSS homepage.